Online publication of court records: circumventing the privacy-transparency trade-off

Tristan Allard, Louis Béziaud, Sébastien Gambs

Online publication of court records: circumventing the privacy-transparency trade-off

Tristan Allard
Univ Rennes, CNRS, IRISA
tristan.allard@irisa.fr

Louis Béziaud
Univ Rennes, CNRS, IRISA
Université du Québec à Montréal
louis.beziaud@irisa.fr

Sébastien Gambs
Université du Québec à Montréal
gambs.sebastien@uqam.ca

ABSTRACT

The open data movement is leading to the massive publishing of court records online, increasing transparency and accessibility of justice, and to the design of legal technologies building on the wealth of legal data available. However, the sensitive nature of legal decisions also raises important privacy issues. Current practices solve the resulting privacy *versus* transparency trade-off by combining access control with (manual or semi-manual) text redaction. In this work, we claim that current practices are insufficient for coping with massive access to legal data (restrictive access control policies is detrimental to openness and to utility while text redaction is unable to provide sound privacy protection) and advocate for a integrative approach that could benefit from the latest developments of the privacy-preserving data publishing domain. We present a thorough analysis of the problem and of the current approaches, and propose a straw man multimodal architecture paving the way to a full-fledged privacy-preserving legal data publishing system.

1 INTRODUCTION

The opening of legal decisions to the public is one of the cornerstones of many modern democracies: it allows to audit and make accountable the legal system by ensuring that justice is rendered with respect to the laws in place. As stated in [11], it can even be considered that *"publicity is the very soul of justice"*. Additionally, in countries following the common law, the access to legal decisions is a necessity as the law in place emerged from the previous decisions of justice courts.

Thus, it is not surprising that the transparency of justice is enshrined in many countries as a fundamental principle, such as the *right to a public hearing* provided by the Article 6 of the European Convention on Human Rights, the Section 135(1) of the Courts of Justice Act (Ontario) stating the general principle that *"all court hearings shall be open to the public"* or in Vancouver Sun (Re) *"The open court principle has long been recognized as a cornerstone of the common law"*. The open data movement push for free access to law with for example the Declaration on Free Access to Law [19]. Multiple open government initiatives also consider the need for an open justice, such as the "Loi pour une République numérique" in France, the Open Government Partnership, the Open Data Charter, the Canada's Action Plan on Open Government. This trend is studied in a report of the OCDE [27], in [49] for the USA or [48] for the UK.

Combined with recent advances in machine learning and natural language processing, the (massive) opening of legal data allows for new practices and applications (called legal technologies). Nonetheless, not all legal decisions should directly be published as such due to the privacy risks that might be incurred by victims, witnesses, members of the jury and judges. Some privacy risks have been considered and mitigated by legal systems for a long time. For instance, the identities of the individuals involved in sensitive cases, such as cases with minors, are usually *anonymized* by default because they belong to a vulnerable subgroup of the population. In situations in which the risks of reprisal are high (*e.g.*, terrorism or organized crimes cases), judges, lawyers and witnesses might also ask for their identities to be hidden [26, 32]. Finally, the identities of the members of a jury are also usually protected to guarantee that they will not be coerced but also to ensure that the strategy deployed by the lawyers is not tailored based on their background. Legal scholars are aware of the need for privacy when opening sensitive legal reports [10, 16, 31].

In the past, these privacy risks were limited due to the efforts that were required to access the decisions themselves. For instance, some countries require to go directly to the court itself to be able to access the legal decisions. Even when the information is available online, the access to legal decisions is usually on a one-to-one basis through a public but restricted API rather than enabling a direct download of the whole legal corpus. Typical restriction mechanisms include CAPTCHAs (SOQUIJ), quotas (CanLII), registration requirement as well as policy agreement and limitation of access to research scholars (Caselaw). Furthermore, the fact that a legal decision is public does not mean that it can, legally, be copied and integrated in other systems or services without any restrictions.

A first approach to limit the privacy risks consists in *redacting* the legal decisions before publishing them. Redaction mostly follows predefined rules that list the information that must be removed or generalized and define how (*e.g.*, by replacing the first and last names by initials, by a pseudonym) [57]. Redaction is in general semi-manual (and sometimes fully manual) because automatic redaction is error-prone [46]. This makes it extremely costly, not scalable, and does not completely remove the risks of errors [57]. For example, 3.9 million decisions are pronounced in France every year but only 180000 are recorded in government databases and less than 15000 are made accessible to the public [28]. Moreover, even a perfect redaction would still offer weak privacy guarantees. A redacted text still contains a non-negligible amount of information, possibly identifying or sensitive, that may be extracted, *e.g.*, from the background of the case or even from the natural language semantics.

Another approach is access control, such as non-publication (*e.g.*, a case involving terrorism was held in secret in Britain [13]), rate limit, or registration requirements. However, access control mechanisms are binary and do not protect the privacy of the texts

for which the access is authorized. Furthermore, restricting massive accesses for blocking also restricts the development of legal technologies that require a massive access to legal data.

In a nutshell, this paper makes the following contributions:

- We clearly state the problem of reconciling transparency with privacy when opening legal data massively.
- We analyze the limits of the current approach, widespread in real-life.
- We propose a high-level straw man architecture of a system for publishing legal data massively in a privacy-preserving manner without precluding the traditional open court principles.

The outline of the paper is as follows. First in Section 2, we state the problem by describing precisely the content of legal data and by explaining the open legal data desiderata. Afterwards, we present the privacy limitations of the current approach, redaction, in Section 3.2, before describing in Section 4 our proposal of architecture for the publication of legal data ensuring both privacy and utility. Finally, we conclude in Section 5.

2 PROBLEM STATEMENT

2.1 Legal data

Legal reports are defined as written documents produced by a court about a particular judgment, which is itself a written decision of a court regarding a particular case (oral judgments are transcribed). Although the content of a case report varies with respect to the court and the country, it can consist of elements such as [56]:

(1) the case name and case citation (identifier);
(2) the date of judgment and the hearing dates;
(3) the court and judges involved in the decision;
(4) the appearances (parties and their representatives);
(5) the statement of facts: identify–sometimes in great length–the relationship and status of the parties, the legally relevant facts (*i.e.*, what happened), and the procedurally significant facts (*e.g.*, cause of action, relief request, raised defenses);
(6) the procedural history: describes–if applicable–the disposition of the case in the lower court(s), the damages awarded, the reason for appeal, etc.;
(7) the issues: point of law in dispute;
(8) the law of the case: elements of law that the court applies;
(9) the concurring and/or dissenting opinions (of judges);
(10) the orders: the decision itself.

We can broadly distinguish three different categories of judicial data depending: metadata, facts and reasoning. Metadata (elements 1, 2, 3 and 4) correspond to identifiers of the case and basic information (*e.g.*, date, parties and judge) and is written mostly in a structured way. Facts (elements 5, 6 and 7) are information pertaining to the parties, disclosing their personal "story". Reasoning (elements 8, 9 and 10) is the logic of the case, which is not specific to the parties.

2.2 Desiderata for the opening of legal data

2.2.1 Need for readability and accessibility. The access to legal decisions is required both for ethical (transparency) and practical reasons such as case law, which is the use of past legal decisions to support the decision for future cases. Thus, the judiciary system is built on the assumption that legal decisions are made public and accessible by default (*open-court principle*), so that (1) citizens are able to inspect decisions as a way to audit the legal system and (2) past decisions can be used to interpret laws, and as such must be known from legal practitioners and citizens. It follows that decisions must be made available in a form readable by humans (*i.e.*, natural language). Natural language format can be opposed to machine-readable formats such as word-vectors representation or logical propositions, which we will discuss later. The need for openness, the current practice in terms of open court, and the associated risks are detailed in [16, 47]. They conclude that, although there are powerful voices in favor of open court, radical changes in access and dissemination require new privacy constraints, and a public debate on the effect of sharing and using information in records.

Accessibility is also an important issue. In the past, the access to decisions required attending public hearings or reading books called "reporters". Later, decisions have started to be shared on digital medium such as compact discs or DVDs for example before being accessible online more recently. For instance, in the USA, CourtListener[1] shares 3.6M decisions and the Caselaw access project[2] 6.7M unique cases; the Canadian Legal Information Institute[3] (CanLII) publishes 2.5M Canadian decisions. The aim of these services is to facilitate access to legal records to individuals–law professionals (judges, lawmakers and lawyers), journalists, or citizens. The online publication also enables the large-scale access and processing of records, in particular due to the standardized format.

2.2.2 Need for massive accesses (legal technologies). The term *legal technologies* broadly encompasses all the technologies used in the context of justice. The website CodeX Techindex[4], a project by the Stanford Center for Legal Informatics, references more than a thousand companies, and defines nine different categories: (1) Marketplace, (2) Document Automation, (3) Practice Management, (4) Legal Research, (5) Legal Education, (6) Online Dispute Resolution, (7) E-Discovery, (8) Analytics and (9) Compliance.

A subset of these categories–2, 4, 7, 8 and 9–requires some form of "understanding" of legal documents, usually performed through natural language processing (NLP) and machine learning (ML) approaches [18, 59]. We focus here on these categories as they are based on the analysis of a large number of legal data. One of the main challenges we have faced is that usually companies provide very few technical details about their actual processing and usage of legal documents.

The automatic processing and analysis of legal records have multiple applications, such as computing similarity between cases [44, 51, 67], predicting legal outcomes [3, 38] (*e.g.*, by weighing the strength of the defender arguments and the legal position of a client in a hypothetical or actual lawsuit), identifying influential cases [45, 53, 63] or important part of laws [52], estimating the risk of recidivism [66], summarizing legal documents [69], extracting entities (*e.g.*, parties, lawyers, law firms, judges, motions, orders, motion type, filer, order type, decision type and judge names) from

[1] https://www.courtlistener.com
[2] https://case.law
[3] https://www.canlii.org
[4] http://techindex.law.stanford.edu

legal documents [17, 60], topic modelling [7, 54], concept mapping [12] or inferring patterns [8, 41].

Focus on text-based legal techs. Most of the technologies introduced in the previous section rely on the processing of large database of legal data. However, the unstructured nature of legal data is one of the main challenges of the application of artificial intelligence in law [2]. Consequently, the analysis of a legal text corpus first requires to apply some pre-processing to add structure to the text. Figure 1 represents an abstract processing pipeline for court files, extracted mostly from academic papers[5], and inferred from the current practice of text analysis and descriptions of associated technologies. In the following, we assume that any application involving the use of machine learning (as highlighted by most legal tech companies) is applied to court records. The first NLP step transforms the unstructured data (*i.e.*, natural language) into some structured representation (see below) by pre-processing it. Afterwards, the second ML step corresponds to the actual application, which is the training (*i.e.*, processing) of the ML algorithm, whose output is represented by the "internal representation" block. The term instance represents the output of the model given some query (*e.g.*, applicable laws given a set of keywords representing infractions).

The pre-processing can be diverse and depends on the task (*e.g.*, extracting a citation graph between cases). However, most NLP-based applications usually rely on a text model. Many models are statistical-based ones, such as document-word-frequency matrix, in which the corpus is decomposed into a matrix in which each cell contains the number of times a particular word appears in a document. This model has multiple variations such as bag-of-words (BoW) [34], term frequency-inverse document frequency, or *n*-grams [71]. For example, a combination of those techniques are used in [3] to predict decisions from the European Court of Human Rights, and by [39] to identify law articles given a query or to answer to questions given a law article. More recent approaches follow a neural network architecture in which a model is trained on the corpus with the objective to predict a word given a context, which is called word embeddings [50]. Multiple variations of this structure exist [36, 40, 42, 43, 72]. This approach has been used for example in [45] to rank and explain influential aspects of law, or by [52] to predict the most relevant sources of law for any given piece of text using "neural networks and deep learning algorithms".

2.2.3 Need for privacy. The massive opening of legal decisions for transparency and technological reasons must not hinder the fundamental rights such as the right to privacy as emphasized by current open justice laws. In particular in this setting, the privacy of a least three main actors must be protected: namely the individuals directly involved in decisions (*i.e.*, the parties), the individuals cited by decisions (*e.g.*, experts or witnesses), and the individuals administering the laws (*i.e.*, magistrates).

However, the problem of publishing legal decisions in a privacy-preserving manner is a difficult one. For instance, authorship attacks [1] may lead to the re-identification of magistrates behind written decisions, or the presence of *quasi-identifiers*[6] within the text decisions may lead to the re-identification of the individuals involved in or cited. Famous real-life examples, such as the governor Weld's [65] or Thelma Arnold's re-identification [6], both based on the exploitation of quasi-identifiers, are early demonstrations of the failure of naive privacy-preserving data publishing schemes. Thus despite the fact that legal decisions are written as unstructured text, structured information can be extracted from them, including the formal argument, the decision itself (*e.g.*, "guilty" or "innocent"), as well as arbitrary information about the individuals involved (*e.g.*, gender, age and social relationships).

Pseudonymization schemes simply consist in removing directly identifying data (*e.g.*, social security number, first name and last name, address) and keeping unchanged the rest of the information (quasi-identifiers included). These schemes provide a very weak protection level, as acknowledged in privacy legislations (*e.g.*, GDPR), which has led to the development of new approaches for sanitizing personal data in the last two decades (see for instance the survey in [14]). In this paper, we focus on privacy-preserving data publishing schemes providing formal privacy guarantees that hold against several publications (as required by any real-life privacy-preserving data publishing system). These schemes are based on (1) *a formal model* stating the privacy guarantees the scheme as well as *a privacy parameter* for tuning the "privacy level" that must be achieved, and (2) *a sanitization algorithm* designed to achieve the chosen model.

A formal model exhibits a set of *composability properties* that defines formally the impact on the overall privacy guarantees of using the scheme on a *log of publications* (also called *disclosures log* in the following). In particular, we will consider the ϵ-differential privacy model [21], defined formally in Definition 2.1, parametrized by ϵ, and achievable by the Laplace mechanism. Its self-composability properties are stated in Theorem 2.2 and its overall privacy guarantees are quantified by the evolution of the disclosures log, and in particular by the evolution of the ϵ value along the various differentially-private releases.

Definition 2.1 (ϵ-differential privacy [21]). The randomized function $\mathtt{f}$ satisfies ϵ-differential privacy, in which $\epsilon > 0$, if:

$$\Pr[\mathtt{f}(\mathcal{D}_1) = O] \leq e^{\epsilon} \cdot \Pr[\mathtt{f}(\mathcal{D}_2) = O]$$

for any set $O \in \mathit{Range}(\mathtt{f})$ and any tabular dataset $\mathcal{D}_1$ and $\mathcal{D}_2$ that differs in at most one row (in which each row corresponds to a distinct individual).

In a nutshell, ϵ-differential privacy ensures that the presence (or absence) of data of a single individual has a limited impact on the output of the computation, thus limiting the inference that can be done by an adversary about a particular individual based on the observed output.

Theorem 2.2 (Sequential and parallel Composability [23]). *Let $\mathtt{f}_i$ be a set of functions such that each provides ϵ_i-differential privacy. First, the* sequential composability *property of differential privacy states that computing all functions on the same dataset results*

[5] The majority of the legal technologies market consists in commercial applications. They do not give information about their inner working and underlying techniques.

[6] A quasi-identifier is a combination of attributes that are usually unique in the population, thus indirectly identifying an individual. A typical example is the triple `(age, zip code, gender)`.

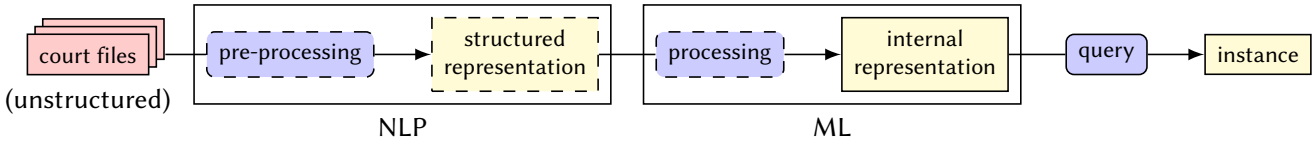

Figure 1: High-level pipeline of court files processing for Legal Techs

in satisfying $(\sum_i \epsilon_i)$*-differential privacy. Second, the* parallel composability *property states that computing each function on disjoint subsets provides* $\max(\epsilon_i)$*-differential privacy.*

3 ANALYSIS OF CURRENT PRACTICES

In the following section, we review the current practice for legal data anonymization and privacy regulations. We also make a connection with medical data anonymization techniques on which most papers rely. To be concrete, we illustrate the privacy risks through examples of re-identification attacks. Finally, we argue that rule-based anonymization is not sufficient to provide a strong privacy protection and discuss the (formal) issues surrounding text anonymization.

3.1 Redaction *in the wild*

Redaction of legal data. The redaction process consists in removing or generalizing a set of predefined terms defined by law through a semi-manual process [57]. Furthermore, access to legal documents or even public hearings can be restricted in well-defined cases. The common practice is to replace sensitive terms, as defined below, by initials, random letters, blanks or generalized terms (*e.g.*, "Montréal" becomes "Québec"). The specific set of rules regarding protected terms and the associated replacement practice can differ between countries and courthouses [57].

According to [58], the following information is to be systematically removed for any person (subject to a restriction on publication), as well as for each of his or her relatives (parents, children, teachers, neighbors, employers, colleagues, school ...):

(1) names,
(2) date and place of birth
(3) contact details (number, street, municipality, postal code, telephone, fax, email, web page, IP address),
(4) unique personal identifiers (social security number, health insurance number, medical file, passport, bank account, credit card, ...),
(5) personal possessions identifiers (license or serial number, cadastral designation, company name, ...)

In some context, the following data is also removed if it can be used to identify one of the individuals aforementioned:

(7) small communities or geographic locations,
(8) the accused and co-accused if their identity is not already protected by law,
(9) the intervenors (court experts, social workers, police officers, doctors, ...)
(10) unusual information (number of children if abnormally high, income if particularly high, exceptional occupation or function).

[16] present numerous examples of legislation framing the publication of specific terms and putting restriction to the *open-court principle*. For instance, it is common by default to hide the identity of victims of sexual offenses or children in youth courts. The identity of jurors and witnesses is also kept secret to avoid coercion or parties tailoring their strategy. In addition in the USA, the fear for national security or the possible prejudice to another trial can lead to a complete ban on reporting being issued.

Paper versus digital. The main difference between paper and digital access is the "practical obscurity" of paper records on the one hand, and the easy accessibility of digital records, on the other. The awkwardness of accessing paper records stored in a public courthouse puts inherent limitations on the ability of individuals or groups to access those records. In contrast, digital records are easy to analyze, can be searched in "bulk" by combining various key factors (*e.g.*, divorce and children) and can potentially be accessed from any computer. Thus, traditional distribution provides "practical obscurity" [15], in that it is inconvenient (*i.e.*, time-consuming) to attend the courthouse or read case reports.

Anonymization of medical data. The Health Insurance Portability and Accountability Act (HIPAA) in the USA defines the security and privacy requirements of health information for both health professionals and technologies involved in medical data. The search for complying with HIPPA has led to an important body of work on the redaction of health records. In particular, automated redaction or generalization of the sensitive terms defined in HIPPA generally involves domain specific named-entity recognition and generalization of terms through medical ontologies. As a concrete anonymization tool, Scrub [64] uses template matching to detect sensitive terms, which are replaced with synthetic data of similar type (*e.g.*, a name with a name, a disease with a similar disease). t-PAT [33] replaces sensitive words or phrases–recognized by an ontology–with more general terms using an early privacy-preserving data publishing model, called k-anonymity [65], to preserve the privacy of patients.

3.2 Limits of current approaches

Our objective in this section is to provide examples of potential attacks in order to illustrate the technical difficulties of raw text anonymization. Figures 3, 2, 4, 5, 6 are excerpts from French and Canadian opinions[7].

A common redaction practice is to replace names by initials as shown in Figure 2. The uniqueness of initials [25] is increased by combining multiple parties, particularly if the relationship between the parties is known (*e.g.*, in a divorce case).

A combination of attributes, which can be extracted using dedicated named-entity recognition, is presented in Figure 3: names of

[7] We translated them using DeepL (https://www.deepl.com)

E.B. Petitioner v. V.I. Respondent

Judgment for Dissolution of Marriage

Figure 2: Droit de la famille – 15334, 2015 QCCS 762 (CanLII), http://canlii.ca/t/ggk9w

Katopodis v. Katopodis

SUPREME COURT OF ONTARIO

The parties were married on August 25, 1968; a daughter was born on November 30, 1972; the parties separated in April, 1977. The wife first went to see Dr. James

Figure 3: Katopodis v. Katopodis, 1979 CanLII 1887 (ON SC), http://canlii.ca/t/g19bb

the association Real Madrid Club de Futbol and several players of this team, Zinedine Z., David B., Raul Gonzalès B. aka Raul, Ronaldo Luiz Nazario de L., aka Ronaldo, and Luis Filipe Madeira C., aka Luis Figo

Figure 4: CA Paris, 11^{e} ch., sect. B, 14 February 2008, *Unibet Ltd c/ Real Madrid et autres*, RG n° 06/11504, GP

the American company Coca Cola Company markets drinks under the French trade mark "Coca Cola light sango", of which it is the proprietor; that M. [...] Abdel X, relying on the infringement of his artist's name and surname, has brought an action for damages against the Coca Cola Company [...] On the ground that Abdel X maintains that, as an author and screenwriter, he is entitled to oppose the use of the name "X" to designate a drink marketed by the companies of the Coca Cola group.

Figure 5: Civ. 1re, 10 April 2013, n° 12-14.525, *Sango c/ Coca-Cola*, D. 2013. 992 ; CCE July 2013, n° 73

parties, parties are divorced, date of marriage, parties have a daughter, birthdate of the daughter, date of divorce, reside in Ontario near a Dr. James. This combination could be used as a quasi-identifier by a re-identification attack.

Figure 4 is anonymized according to the CNIL recommendations of 2006, which requires the last name of individuals to be replaced by its initial. However, widely available background knowledge on the "Real Madrid Club de Futbol" combined with the (real-life) pseudonyms of the "players" trivially leaks their identity.

The de-anonymization of Figure 5 relies on the text semantics instead of background knowledge. It requires the adversary (1) to identify the link (X) between "M. [...] Abdel X" and "the use of the name 'X' to designate a drink", and (2) to infer that the drink is called "sango", thus leading to the conclusion that X = "sango". While this attack may not be easy to automatize due to the hardness of detecting the semantics inference, it is, however, trivial to perform for a human (*e.g.*, by crowdsourcing it).

X, born [...] 2017; Y, born [...] 2018 the children and C; D the parents

Applications are submitted for X, aged 1 year, and Y, aged 2 months. The Director of Youth Protection (DYP) would like X to be entrusted to her aunt, Ms. E, until June 25, 2019. As for Y, that he be entrusted to a foster family for the next nine months. The father has two other children, Z and A, from his previous union with Mrs. F. The mother has another child, B, from her union with Mr. G.

Figure 6: Protection de la jeunesse – 186470, 2018 QCCQ 6920 (SOQUIJ), http://t.soquij.ca/x4L6N

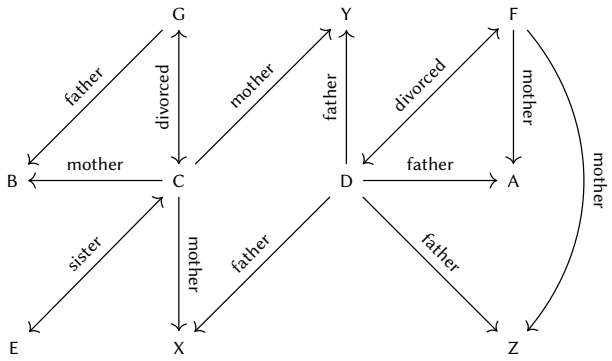

Figure 7: Relationship graph manually extracted from Figure 6

Similar to Figure 3, Figure 6 could be attacked through a combination of attributes and relationship (*e.g.*, extracted with Snorkel [61]). This opinion from the Youth court involves children and, as such, follows the strictest anonymization rules of the SOQUIJ: only the year's birthdate of children is given and names are replaced by random letters. However, an adversary can extract an extensive relationship graph (see Figure 7), which could be matched over a relationship database (*e.g.*, Facebook). In this case, a quasi-identifier could be the relationship graph (or parts of it).

A study [5] has shown that generalization-based sanitization is vulnerable to correlation attacks in which the generalized terms can be correlated with other terms, seemingly non-identifying, in order to jeopardize the effects of the generalization and consequently disclose sensitive terms.

Besides the content of legal documents, stylometry [55] can be used to identify authors (*i.e.*, magistrates) by their writing style. Mitigation for this kind of attack exist [24, 68] but their output is only machine readable. Similarly, it is possible to exploit decision patterns to re-identify judges, as done for the Supreme Court of the United States [38].

3.3 Reasons for the failure of rule-based redaction

The review of current practices for tackling the privacy of legal documents in Section 3.1 has highlighted the widespread use of

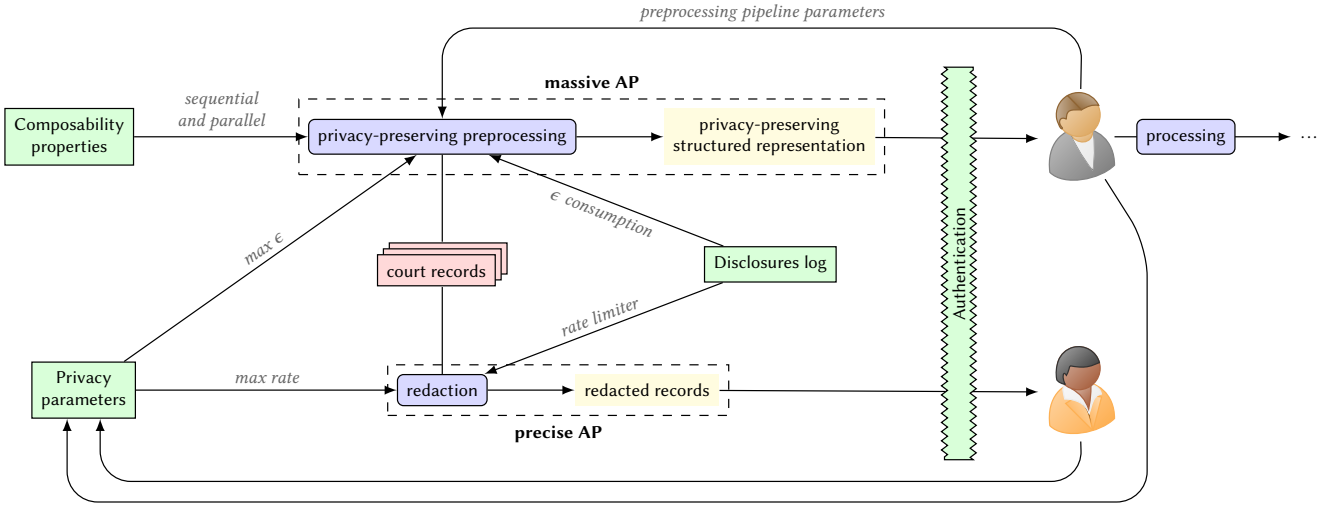

Figure 8: Multimodal publication architecture

rule-based redaction, in which a set of patterns is defined as being sensitive and is either removed or replaced. However, as shown in Section 3.2 (1) privacy can be violated even in "simple" instances and (2) identifying information remains in most cases. In other words, rule-based redaction does not provide any sound privacy guarantee. We observe that it suffers from the following main difficulties.

(1) *Missing rule difficulty.* Many combinations of quasi-identifiers can lead to re-identification and the richness of the output space offered by natural language (*i.e.*, what can be expressed) can hardly be constrained to a set of rules. Furthermore, identifying the sensitive terms is challenging and domain-specific. This issue is the subject of multiple studies in the context of medical data [9, 20, 64].
(2) *Missing match difficulty.* The current state of the art about relationship extraction and named-entity recognition makes it hard to ensure that all terms that should be redacted will be detected, in particular because of the many possible ways to express the same idea (*e.g.*, *circumlocution*).

Although these observations make the rule-based redaction difficult, it is important to note that attacks, *e.g.*, re-identification, remain simpler than protection. Indeed, an adversary has to find a single attack vector (*i.e.*, a missing rule or a missing pattern) whereas the redaction process needs to consider all the possibilities.

4 MULTIMODAL PUBLICATION SCHEME

In Section 2, we have shown that the publication of legal documents serves two distinct and complementary purposes: (1) the traditional objective of transparency and case law, and (2) the modern objective of legal technologies of providing services to citizens and legal professionals. These two purposes obey to different utility and privacy requirements. More precisely, the traditional use case requires human-readable documents while legal techs need a machine-readable format for automated processing. Moreover, transparency and case law involve the access to opinions on an individual basis (*i.e.*, one-at-a-time), similarly to attending a hearing in person. In contrast, legal technologies rely on the access to massive legal databases. This difference in cardinality (*i.e.*, one versus many) entails different privacy risks. In particular, the massive processing of legal data requires the use of a formal privacy framework with composability properties (see Section 2.2.3). All this suggests the inadequacy of any *one-size-fits-all* approach.

4.1 Access modes

As a consequence, we propose that the organization in charge of the publication of legal decisions should consider two modes of publication: the *precise access mode* and the *massive access mode.*

Precise access mode. To fulfill the "traditional" use case, the precise access mode provides full access to legal decisions that are only redacted using the current practices. This access mode is designed for the transparency and case law usages, and is to be used typically by individuals (*e.g.*, law professionals, journalists and citizens). Similar to the "traditional" paper-based publication scheme, in the precise access mode [29], a user has access to text documents, either in full or only extracts (partial access is useful for crowdsourcing tagging in order to build a dataset). While the current practice of redacting identifiers could be combined with more automated approaches such as [30, 62]. The aim of this mode is to provide strong utility first. It is similar to the websites currently publishing legal documents (*e.g.*, Legifranceor CanLII), as it allows browsing, searching and reading documents.

To prevent malicious users from diverting the precise access mode for performing massive accesses, users must be authenticated and their access must be restricted (*e.g.*, rate limitation or proof of work [22]). The main objective of the restricted access is to make it difficult to rebuild the full (massive) database. In addition, this mode provides privacy through "practical obscurity" similarly to the paper-based system.

Massive access mode. The massive access mode gives access only to pre-processed data resulting from privacy-preserving versions of the standard NLP pipelines available on the server, *i.e.*, aggregated

and structured data extracted from or computed over large numbers of decisions, as required for the "modern" use case. It should be compatible with most legal tech applications that traditionally use a database of legal documents (see Section 2.2.2). Note that the perturbations due to privacy-preserving data publishing schemes have usually less impact (in terms of information loss) when applied late in the pipeline (see Figure 8), at the cost of a loss of generality of the output.

Users need to be able to tune the pre-processing applied. For the sake of simplicity, we assume that the user (*i.e.*, legaltech developer) provides the parameters for a given NLP pipeline (see Fig. 8). These parameters can be for instance the maximum number of features or n-grams range to consider in the case of a BoW model or the window for word embeddings. But more complex implementations can be designed, *e.g.*, allowing experiments by the users, fine-tuning for each dataset/task, as well as customization (*e.g.*, for cleaning the data). This can be done (1) by generating structured synthetic *testing* data (*e.g.*, a set of features extracted from legal data) in a privacy-preserving manner (*e.g.*, PATE-GAN [35]) and (2) by designing a full pre-processing pipeline that embeds privacy-preserving calls to the server (*e.g.*, through a privacy-preserving computation framework such as Ektelo [70]).

The massive access mode must also authenticate users in order to monitor the overall privacy guarantees satisfied for each user based on his disclosures log and on the composability properties of the privacy-preserving data publishing schemes used.

As a result, the data is protected using authentication and strong privacy definitions as presented in Section 2.2.3. Examples of applications of differential privacy to NLP models include [24], which adds noise to word-frequency-matrix to achieve differential privacy, or [68], which samples the dictionary of the model using the differentially-private exponential mechanism [23]. The aim of these two approaches is to protect against authorship attribution.

[24] uses a relaxation of differential privacy, d_χ-privacy [4] which allows the authors to consider a distance between documents computed using word embeddings, rather than the row-based distance presented in Definition 2.1. The objective is to modify BoW representation of documents "similar in topics" remain "similar to each other" (w.r.t. the metrics defined on word embeddings), irrespective of authorship. In practice, this is achieved by drawing BoW where the probability of each word being associated to a document is distributed according to a Laplace probability density function.

The goal of [68] is to derive a differentially private synthetic feature vectors, keeping the theme of each document while preventing authorship attribution. Feature vectors map a set of words (the dictionary) to probabilities of the word appearing in each document. The main idea of the approach is to sample the dictionary from a reference dictionary (*e.g.*, using synonyms from WordNet's synsets) using the differentially private exponential mechanism.

In practice, the massive access mode can be plugged into the existing platforms that store massive number of legal documents and already support the precise access mode, such as CourtListener or CanLII.

Finally, another potential need is the annotation of documents, which is the addition to terms, sentences, paragraphs or documents of metadata such as syntax information (*e.g.*, verb or noun), semantic, pragmatic (*e.g.*, presupposition and implicature). This step is crucial in NLP, and is usually done manually, for example through crowdsourcing. Crowdsourcing-specific approaches for privacy-preserving task processing [37] require to split the task (*i.e.*, annotation of a set of documents) between non-colluding workers (*e.g.*, at the sentence level) before aggregating the result. Such approach is compatible with our architecture assuming the aggregation is done locally on the platform.

4.2 System overview

We now outline an abstract architecture for a privacy-preserving data publishing system for legal decisions. Our objective is not to provide exhaustive implementation guidelines, but rather to identify the key components that such an architecture should possess.

Figure 8 depicts the proposed architecture. The precise and massive access modes are both protected by the `Authentication` module. The `Authentication` module can be implemented by usual strong authentication techniques (*e.g.*, for preventing impersonation attacks). Authentication is necessary for enforcing the access control policy through the `Access Control` module and for maintaining for each user his `Disclosure Log`. The log contains all the successful access requests performed by a user. It is required for verifying that the overall privacy guarantees are not breached, *e.g.*, the rate limitation is not exceeded for the precise access mode, or the composition of the privacy-preserving data publishing schemes, formalized in the `Composability Properties`, does not exceed the disclosure allowed. Finally, the `Privacy Parameters` contain the overall privacy guarantees that must always hold, defined by the administrator (*e.g.*, rate limit or higher bound on the ϵ differential privacy parameter). The user may additionally be allowed to tune the privacy parameters input by a privacy-preserving data publishing scheme (*e.g.*, the fraction spent in the higher bound on the ϵ differential privacy parameter) provided it does not jeopardize the overall privacy guarantees.

5 DISCUSSION

In this paper, we analyzed the needs for publishing legal data and the limitations of rule-based redaction (*i.e.*, the current approach) for fulfilling them successfully. We proposed to discard any one-size-fits-all approach and outlined a straw man architecture balancing the utility and privacy requirements by distinguishing the traditional, one-to-one, use of legal data from the modern, massive, use of legal data by legal technologies. Our proposition can easily be implemented on current platforms.

6 ACKNOWLEDGMENTS

This work was partially funded by the PROFILE-INT project funded by the LabEx CominLabs (ANR-10-LABX-07-01). Sébastien Gambs is supported by the Canada Research Chair program as well as by a Discovery Grant from NSERC and the Legalia project from the AUDACE program funded by the FQRNT.

REFERENCES

[1] Ahmed Abbasi and Hsinchun Chen. Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Transactions on Information Systems (TOIS)*, 26(2):7, 2008. doi: 10.1145/1344411.1344413.

[2] Benjamin Alarie, Anthony Niblett, and Albert H Yoon. How artificial intelligence will affect the practice of law. *University of Toronto Law Journal*, 68(supplement 1):106–124, 2018.

[3] Nikolaos Aletras, Dimitrios Tsarapatsanis, Daniel Preoţiuc-Pietro, and Vasileios Lampos. Predicting judicial decisions of the european court of human rights: A natural language processing perspective. *PeerJ Computer Science*, 2016. doi: 10.7717/peerj-cs.93.

[4] Mário S Alvim, Konstantinos Chatzikokolakis, Catuscia Palamidessi, and Anna Pazii. Metric-based local differential privacy for statistical applications. *arXiv preprint arXiv:1805.01456*, 2018.

[5] Balamurugan Anandan and Chris Clifton. Significance of term relationships on anonymization. In *Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology-Volume 03*, pages 253–256. IEEE Computer Society, 2011.

[6] Michael Arrington. AOL Proudly Releases Massive Amounts of Private Data. *TechCrunch*, 2006. URL https://social.techcrunch.com/2006/08/06/aol-proudly-releases-massive-amounts-of-user-search-data/.

[7] Kevin D. Ashley and Stefanie Brüninghaus. Automatically classifying case texts and predicting outcomes. *Artif. Intell. Law*, 17(2):125–165, June 2009. ISSN 0924-8463. doi: 10.1007/s10506-009-9077-9.

[8] Kevin D Ashley and Vern R Walker. Toward constructing evidence-based legal arguments using legal decision documents and machine learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law*, pages 176–180, 2013.

[9] Mikhail J Atallah, Craig J McDonough, Victor Raskin, and Sergei Nirenburg. Natural language processing for information assurance and security: an overview and implementations. In *Proceedings of the 2000 workshop on New security paradigms*, pages 51–65, 2001.

[10] Jane Bailey and Jacquelyn Burkell. Revisiting the open court principle in an era of online publication: Questioning presumptive public access to parties' and witnesses' personal information. *Ottawa L. Rev.*, 48:143, 2016.

[11] Jeremy Bentham and John Bowring. *The Works of Jeremy Bentham*, volume 4. W. Tait, 1843.

[12] Stefanie Brüninghaus and Kevin D. Ashley. Using machine learning for assigning indices to textual cases. In David B. Leake and Enric Plaza, editors, *Case-Based Reasoning Research and Development*, pages 303–314, Berlin, Heidelberg, 1997. Springer Berlin Heidelberg. ISBN 978-3-540-69238-6.

[13] Krishnadev Calamur. In a first for Britain, a secret trial for terrorism suspects. *NPR*, 2014. URL https://text.npr.org/s.php?sId=319076959.

[14] Bee-Chung Chen, Daniel Kifer, Kristen LeFevre, and Ashwin Machanavajjhala. Privacy-preserving data publishing. *Found. Trends Databases*, 2(1–2):1–167, January 2009. ISSN 1931-7883. doi: 10.1561/1900000008.

[15] Judges Technology Advisory Committee. Open courts, electronic access to court records, and privacy: discussion paper. Technical report, Canadian Judicial Council, 2003. URL http://publications.gc.ca/collections/collection_2008/lcc-ccdc/JL2-75-2003E.pdf.

[16] Amanda Conley, Anupam Datta, Helen Nissenbaum, and Divya Sharma. Sustaining privacy and open justice in the transition to online court records: A multidisciplinary inquiry. *Md. L. Rev.*, 71:772, 2011.

[17] Tonya Custis, Frank Schilder, Thomas Vacek, Gayle McElvain, and Hector Martinez Alonso. Westlaw Edge AI Features Demo: KeyCite Overruling Risk, Litigation Analytics, and WestSearch Plus. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law - ICAIL '19*, pages 256–257, Montreal, QC, Canada, 2019. ACM Press. ISBN 978-1-4503-6754-7. doi: 10.1145/3322640.3326739. URL http://dl.acm.org/citation.cfm?doid=3322640.3326739.

[18] Robert Dale. Law and word order: NLP in legal tech. *Natural Language Engineering*, 25(1):211–217, January 2019. doi: 10.1017/s1351324918000475. URL https://www.cambridge.org/core/product/identifier/S1351324918000475/type/journal%5Particle.

[19] Declaration on Free Access to Law. Declaration on free access to law, 2002. URL http://www.worldlii.org/worldlii/declaration/.

[20] MM Douglass, GD Cliffford, Andrew Reisner, WJ Long, GB Moody, and RG Mark. De-identification algorithm for free-text nursing notes. In *Computers in Cardiology, 2005*, pages 331–334. IEEE, 2005.

[21] Cynthia Dwork. Differential privacy. In *Proceedings of the 33rd International Conference on Automata, Languages and Programming - Volume Part II*, volume 4052 of *Icalp'06*, pages 1–12, Berlin, Heidelberg, July 2006. Springer-Verlag. ISBN 3-540-35907-9, 978-3-540-35907-4. doi: 10.1007/11787006_1. URL https://www.microsoft.com/en-us/research/publication/differential-privacy/.

[22] Cynthia Dwork and Moni Naor. Pricing via processing or combatting junk mail. In *Annual International Cryptology Conference*, pages 139–147. Springer, 1992.

[23] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[24] Natasha Fernandes, Mark Dras, and Annabelle McIver. Generalised differential privacy for text document processing. In *International Conference on Principles of Security and Trust*, pages 123–148. Springer, 2019. doi: 10.1007/978-3-030-17138-4_6.

[25] Craig Finseth. Rfc 1439 uniqueness of unique identifiers. RFC 1439, RFC Editor, March 1993.

[26] Caroline Fleuriot. Avec l'accès gratuit à toute la jurisprudence, des magistrats réclament l'anonymat. *Dalloz Actualité*, February 2017. URL https://www.dalloz-actualite.fr/flash/avec-l-acces-gratuit-toute-jurisprudence-des-magistrats-reclament-l-anonymat.

[27] Organisation for Economic Co-operation and Development, editors. *The call for innovative and open government: an overview of country initiatives*. Oecd, Paris, 2011. ISBN 9789264107045.

[28] Amaury Fouret, Mathieu Perez, Valentin Barrière, Edouard Rottier, and Éloi Buat-Ménard. Open Justice. Technical report, Cour de cassation, 2019. URL https://entrepreneur-interet-general.etalab.gouv.fr/defis/2019/openjustice.html.

[29] Woodrow Hartzog and Frederic Stutzman. The case for online obscurity. *Calif. L. Rev.*, 101:1, 2013.

[30] Fadi Hassan, David Sánchez, Jordi Soria-Comas, and Josep Domingo-Ferrer. Automatic Anonymization of Textual Documents: Detecting Sensitive Information via Word Embeddings. In *2019 18th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/13th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE)*, pages 358–365, 2019. doi: 10.1109/TrustCom/BigDataSE.2019.00055.

[31] Joseph Jaconelli. *Open Justice: A critique of the public trial*. Oxford University Press on Demand, 2002.

[32] Jean-Baptiste Jacquin. Terrorisme : la peur des magistrats. *Le Monde*, January 2017. URL https://www.lemonde.fr/police-justice/article/2017/01/19/terrorisme-la-peur-des-magistrats_5065242_1653578.html.

[33] Wei Jiang, Mummoorthy Murugesan, Chris Clifton, and Luo Si. t-plausibility: Semantic preserving text sanitization. In *2009 International Conference on Computational Science and Engineering*, volume 3, pages 68–75. IEEE, 2009.

[34] Thorsten Joachims. Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning*, pages 137–142. Springer, 1998. doi: 10.1007/bfb0026683.

[35] James Jordon, Jinsung Yoon, and Mihaela van der Schaar. PATE-GAN: generating synthetic data with differential privacy guarantees. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=S1zk9iRqF7.

[36] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*, 2016. doi: 10.18653/v1/e17-2068.

[37] Hiroshi Kajino, Yukino Baba, and Hisashi Kashima. Instance-privacy preserving crowdsourcing. In *Second AAAI Conference on Human Computation and Crowdsourcing*, 2014.

[38] Daniel Martin Katz, Michael J. Bommarito II, and Josh Blackman. A general approach for predicting the behavior of the supreme court of the united states. *PLoS One*, 12(4), 2017. doi: 10.1371/journal.pone.0174698.

[39] Mi-Young Kim, Juliano Rabelo, and Randy Goebel. Statute Law Information Retrieval and Entailment. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law - ICAIL '19*, pages 283–289, Montreal, QC, Canada, 2019. ACM Press. ISBN 978-1-4503-6754-7. doi: 10.1145/3322640.3326742. URL http://dl.acm.org/citation.cfm?doid=3322640.3326742.

[40] Yoon Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014. doi: 10.3115/v1/d14-1181.

[41] Fred Kort. Quantitative analysis of fact-patterns in cases and their impact on judicial decisions. *Harv. L. Rev.*, 79:1595, 1965.

[42] Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. Recurrent convolutional neural networks for text classification. In *Twenty-ninth AAAI conference on artificial intelligence*, 2015.

[43] Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. Recurrent neural network for text classification with multi-task learning. *arXiv preprint arXiv:1605.05101*, 2016.

[44] Arpan Mandal, Raktim Chaki, Sarbajit Saha, Kripabandhu Ghosh, Arindam Pal, and Saptarshi Ghosh. Measuring similarity among legal court case documents. In *Proceedings of the 10th Annual ACM India Compute Conference*, Compute '17, pages 1–9, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450353236. doi: 10.1145/3140107.3140119.

[45] Max RS Marques, Tommaso Bianco, Maxime Roodnejad, Thomas Baduel, and Claude Berrou. Machine learning for explaining and ranking the most influential matters of law. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*, pages 239–243. Acm, 2019.

[46] Mónica Marrero, Julián Urbano, Sonia Sánchez-Cuadrado, Jorge Morato, and Juan Miguel Gómez-Berbís. Named entity recognition: fallacies, challenges and opportunities. *Computer Standards & Interfaces*, 35(5):482–489, 2013.

[47] Peter W Martin. Online access to court records-from documents to data, particulars to patterns. *Vill. L. Rev.*, 53:855, 2008.

[48] Tom McClean. Not with a bang but a whimper: The politics of accountability and open data in the uk. In *APSA 2011 Annual Meeting Paper*, 2011.

[49] Patrice McDermott. Building open government. *Government Information Quarterly*, 27(4):401–413, 2010.

[50] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[51] Akshay Minocha and Navjyoti Singh. Legal document similarity using triples extracted from unstructured text. In Georg Rehm, Víctor Rodríguez-Doncel, and Julián Moreno-Schneider, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France, May 2018. European Language Resources Association (ELRA). ISBN 979-10-95546-18-4.

[52] Ivan Mokanov, Daniel Shane, and Benjamin Cerat. Facts2law: using deep learning to provide a legal qualification to a set of facts. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*, pages 268–269. Acm, 2019.

[53] Martin Možina, Jure Žabkar, Trevor Bench-Capon, and Ivan Bratko. Argument based machine learning applied to law. *Artificial Intelligence and Law*, 13(1): 53–73, 2005.

[54] Ramesh Nallapati and Christopher D Manning. Legal docket classification: Where machine learning stumbles. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 438–446, 2008.

[55] Tempestt Neal, Kalaivani Sundararajan, Aneez Fatima, Yiming Yan, Yingfei Xiang, and Damon Woodard. Surveying stylometry techniques and applications. *ACM Computing Surveys (CSUR)*, 50(6):1–36, 2017.

[56] University of Houston Law Center. How to brief a case. Technical report, University of Houston Law Center, 2009. URL https://www.law.uh.edu/lss/casebrief.pdf.

[57] Marc Opijnen, Ginevra Peruginelli, Eleni Kefali, and Monica Palmirani. On-line publication of court decisions in the EU: Report of the policy group of the project "building on the european case law identifier". *Available at SSRN 3088495*, 2017.

[58] Luc Plamondon, Guy Lapalme, and Frédéric Pelletier. Anonymisation de décisions de justice. In *XIe Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2004)*., pages 367–376, Fès, Maroc, May 2004. Bernard Bel et Isabelle Martin. (éditeurs), Bernard Bel et Isabelle Martin. (éditeurs). URL http://rali.iro.umontreal.ca/rali/sites/default/files/publis/0UdeM-taln-04.pdf.

[59] Sabrina Praduroux, Valeria de Paiva, and Luigi di Caro. Legal tech start-ups: State of the art and trends. In *Proceedings of the Workshop on MIning and REasoning with Legal texts collocated at the 29th International Conference on Legal Knowledge and Information Systems*, 2016.

[60] Paulo Quaresma and Teresa Gonçalves. Using linguistic information and machine learning techniques to identify entities from juridical documents. In Enrico Francesconi, Simonetta Montemagni, Wim Peters, and Daniela Tiscornia, editors, *Semantic Processing of Legal Texts: Where the Language of Law Meets the Law of Language*, pages 44–59. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010. ISBN 978-3-642-12837-0. doi: 10.1007/978-3-642-12837-0_3.

[61] Alexander Ratner, Stephen H Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. Snorkel: Rapid training data creation with weak supervision. *The VLDB Journal*, 29(2):709–730, 2020.

[62] David Sanchez, Montserrat Batet, and Alexandre Viejo. Automatic general-purpose sanitization of textual documents. *IEEE Transactions on Information Forensics and Security*, 8(6):853–862, 2013. doi: 10.1109/tifs.2013.2239641.

[63] Daniel J Siegel. Cara: An assistance to help find the cases you missed. *Law Prac.*, 43:22, 2017.

[64] Latanya Sweeney. Replacing personally-identifying information in medical records, the scrub system. In *Proceedings of the AMIA annual fall symposium*, page 333. American Medical Informatics Association, 1996.

[65] Latanya Sweeney. K-anonymity: A model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, 10(5):557–570, October 2002. ISSN 0218-4885. doi: 10.1142/S0218488502001648.

[66] Sarah Tan, Julius Adebayo, Kori Inkpen, and Ece Kamar. Investigating human+ machine complementarity for recidivism predictions. *arXiv preprint arXiv:1808.09123*, 2018.

[67] D Thenmozhi, Kawshik Kannan, and Chandrabose Aravindan. A text similarity approach for precedence retrieval from legal documents. In *FIRE (Working Notes)*, pages 90–91, 2017.

[68] Benjamin Weggenmann and Florian Kerschbaum. Syntf: Synthetic and differentially private term frequency vectors for privacy-preserving text mining. *arXiv preprint arXiv:1805.00904*, 2018. doi: 10.1145/3209978.3210008.

[69] Mehdi Yousfi-Monod, Atefeh Farzindar, and Guy Lapalme. Supervised machine learning for summarizing legal documents. In Atefeh Farzindar and Vlado Kešelj, editors, *Advances in Artificial Intelligence*, pages 51–62, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg. ISBN 978-3-642-13059-5.

[70] Dan Zhang, Ryan McKenna, Ios Kotsogiannis, Michael Hay, Ashwin Machanavajjhala, and Gerome Miklau. Ektelo: A framework for defining differentially-private computations. In *Proceedings of the 2018 International Conference on Management of Data*, SIGMOD '18, pages 115–130, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450347037. doi: 10.1145/3183713.3196921.

[71] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, pages 649–657, 2015.

[72] Jianming Zheng, Yupu Guo, Chong Feng, and Honghui Chen. A hierarchical neural-network-based document representation approach for text classification. *Mathematical Problems in Engineering*, 2018, 2018. doi: 10.1155/2018/7987691.